Ein Ebenenmodell für die semantische Integration von Primärdaten und Publikationen in Digitalen Bibliotheken

Stempfhuber, Maximilian; Zapilko, Benjamin

Postprint / Postprint
Sammelwerksbeitrag / collection article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Stempfhuber, M., & Zapilko, B. (2013). Ein Ebenenmodell für die semantische Integration von Primärdaten und Publikationen in Digitalen Bibliotheken. In H. P. Ohly (Hrsg.), *Wissen - Wissenschaft - Organisation: Proceedings der 12. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation* (S. 1-11). Würzburg: Ergon-Verl. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-46473-2

Ein Ebenenmodell für die semantische Integration von Primärdaten und Publikationen in Digitalen Bibliotheken

Maximilian Stempfhuber
RWTH Aachen, Hochschulbibliothek,
52062 Aachen
Tel.: 0241 – 80 - 94459
stempfhuber@bth.rwth-aachen.de

Benjamin Zapilko
GESIS – Leibniz-Institut für Sozialwissenschaften,
Informationelle Prozesse in den Sozialwissenschaften
Lennéstr. 30, 53113 Bonn
Tel.: 0228 – 2281 - 0
benjamin.zapilko@gesis.org

Keywords:

Text-Fakten-Integration; Semantische Integration; Semantische Heterogenitätsbehandlung; Digitale Bibliotheken; Ontologien

Zusammenfassung:

Digitale Bibliotheken stehen derzeit vor der Herausforderung, den veränderten Informationsbedürfnissen ihrer wissenschaftlichen Nutzer nachzukommen und einen integrierten Zugriff auf verschiedene Informationsarten (z.B. Publikationen, Primärdaten, Wissenschaftler- und Organisationsprofile, Forschungsprojektinformationen) zu bieten, die in zunehmenden Maße digital zur Verfügung stehen und diese in virtuellen Forschungsumgebungen verfügbar zu machen. Die daraus resultierende Herausforderungen struktureller und semantischer Heterogenität werden durch ein weites Feld von verschiedenen Metadaten-Standards, Inhaltserschließungsverfahren sowie Indexierungsansätze für verschiedene Arten von Information getragen. Bisher existiert jedoch kein allgemeingültiges, integrierendes Modell für Organisation und Retrieval von Wissen in Digitalen Bibliotheken.

Dieser Beitrag stellt aktuelle Forschungsentwicklungen und -aktivitäten vor, die die Problematik der semantischen Interoperabilität in Digitalen Bibliotheken behandeln und präsentiert ein Modell für eine integrierte Suche in textuellen Daten (z.B. Publikationen) und Faktendaten (z.B. Primärdaten), das verschiedene Ansätze der aktuellen Forschung aufgreift und miteinander in Bezug setzt. Eingebettet in den Forschungszyklus treffen dabei traditionelle Inhaltserschließungsverfahren für Publikationen auf neuere ontologie-basierte Ansätze, die für die Repräsentation komplexerer Informationen und Zusammenhänge (z.B. in sozialwis-

senschaftlichen Umfragedaten) geeigneter scheinen. Die Vorteile des Modells sind (1) die einfache Wiederverwendbarkeit bestehender Wissensorganisationssysteme sowie (2) ein geringer Aufwand bei der Konzeptmodellierung durch Ontologien.

1 Einleitung

In den vergangenen Jahren lässt sich bei Digitalen Bibliotheken ein deutlicher Wandel in ihrer Rolle und Aufgabe für wissenschaftliche Nutzer beobachten (vgl. DELOS 2005b; Gold 2007). Ergebnisse zahlreicher Studien (vgl. Poll 2004) belegen, dass ein Harvesting und Verlinken von Metadaten verschiedener Informationsquellen und deren Aufbereitung für eine gemeinsame Recherche durch die Anwendung von Standardisierungstechniken und -verfahren die Bedürfnisse der Nutzer nicht mehr befriedigen. Eine enge Integration verschiedener Informationsarten (Volltexte, Literaturnachweise, Umfrage- und andere statistische Daten, Zeitreihen, Projektinformationen, etc.) wird immer häufiger erwartet und vorausgesetzt. Diese Erwartungshaltung von Nutzerseite reflektiert die unterschiedliche Verwendung dieser Informationsarten zu verschiedenen Stadien des Forschungsprozesses. Beispielsweise werden in einer frühen Phase der Planung eher Publikationen und Projektinformationen gesucht, während Wissenschaftler in späteren Stadien ihrer Forschungsarbeit z.B. eher Forschungsdaten benötigen, um Sekundäranalysen durchzuführen oder ihre Ergebnisse zu verifizieren.

Besonders in den Sozialwissenschaften, bei denen einerseits Datenarchive, die empirische Daten auf einem sehr detaillierten Level dokumentieren, auf internationaler Ebene organisiert und vernetzt sind, sind diese Informationen und Infrastrukturen andererseits kaum oder nur minimal mit entsprechenden Dokumentbeständen von Bibliotheken oder Informationszentren verbunden. Die Herausforderung für Informationsanbieter liegt daher nicht nur im Aufbau und in der Organisation einer Kollaboration, um diese Informationsbestände zusammen zu führen, sondern auch in den Fragestellungen, wie diese heterogenen Forschungsinformationen auf technischer, struktureller und semantischer Ebene integriert werden können. Die Komplexität, die bei der Dokumentation des vollständigen Datenlebenszyklus einhergeht, entsteht beispielsweise bei verschiedenen Versionen von Fragebögen oder Datensätzen, den dazugehörigen Codebüchern und Variablen, etc. und führt zu sehr speziellen, domänenspezifischen semantischen Strukturen, die derzeit nicht zufriedenstellend auf die semantische Repräsentation von beispielsweise Forschungsliteratur abgebildet werden können.

Das seit einigen Jahren aufkommende e-Science-Paradigma (vgl. Gold 2007), das vor allem als „enhanced" Science verstanden wird, setzt den Fokus auf die Entwicklung und Etablierung einer Infrastruktur bestehend aus Hard- und Software sowie (kollaborativen) Netzwerken, um wissenschaftliche Arbeitsabläufe zu unterstützen und zu erweitern, angefangen bei der Datenakquise bis hin zu neuen wissenschaftlichen Publikationsformen (z.B. elektronisches Publizieren, Open Access Repositorien). Gleichzeitig sollen alle Forschungsergebnisse frei für Recherche und Zugriff zur Verfügung stehen. Für dieses Vorhaben werden wissenschaftliche Modelle und Methoden benötigt, die alle Arten von Forschungsinformationen sowohl strukturell als auch semantisch einheitlich ausdrücken und weiter verarbeiten können sowie Abbildungen definieren und anwenden können, um verwandte Informationen zu identifizieren, miteinander zu verbinden und schließlich für die Nutzer auszuzeichnen. Diese

Anforderungen lassen sich nicht nur bei Bibliotheken erkennen, sondern auch verstärkt auf Seite der Datenarchive (vgl. ARL 2006).

2 Aktuelle Forschungsaktivitäten

In vielen Domänen und Organisationen finden bereits, auch auf internationaler Ebene, Bestrebungen statt, sich mit der derzeitigen Situation und den beschriebenen Beobachtungen auseinanderzusetzen. Ansätze und Aktivitäten lassen sich hierbei auf verschiedensten Abstraktionsebenen identifizieren, beginnend auf den Gebieten der Informationsarchitektur und der formalen Organisation von Digitalen Bibliothekssystemen bis hin zur strukturellen und semantischen Ebene, bei denen ein Umgang mit der Interoperabilität zwischen heterogenen und verteilten Informationsarten einen relevanten Forschungsaspekt einnimmt. Viele dieser Entwicklungen sind jedoch nur wenig oder gar nicht miteinander verbunden und stellen nur selten einen Bezug zu Entwicklungen auf anderen Abstraktionsebenen her. Gleichzeitig wird die Nutzung und Anwendung von Technologien und Standards des Semantic Web für Digitale Bibliotheken zu, z.B. die Verwendung von Ontologien, semantische Annotationen oder deduktive Verfahren diskutiert (vgl. Sure & Studer 2005; Goble et al. 2006; Svensson 2007).

Das DELOS, Network of Excellence on Digital Libraries, verfasste einen State-of-the-Art-Report (DELOS 2005a) über semantische Interoperabilität in Digitalen Bibliotheken. Darin wird ein breites Spektrum an Forschungsaktivitäten beschrieben. Die folgenden Abschnitte geben einen Überblick über derzeitige Aktivitäten und Entwicklungen im Feld der semantischen Interoperabilität.

2.1 Architektonischer und organisatorischer Kontext

Formale Modelle, wie beispielsweise das 5S-Modell (vgl. Goncalves et al. 2004), und Referenzarchitekturen (z.B. Candela et al. 2006) wurden entwickelt, um Digitale Bibliotheken als Ganzes zu strukturieren, zu organisieren und damit eine hohe Effektivität zu erreichen und den Ansprüchen der Nutzer gerecht zu werden. Diese Architekturen berücksichtigen nahezu jeden Aspekt, der für eine effiziente Digitale Bibliothek in einer durch Netzwerke verbundenen Welt relevant ist. Bei solch einem weiten und damit zwangsläufig etwas distanzierten Blickwinkel auf die Gesamtproblematik können Detailprobleme, die auf tieferliegenden Abstraktionsebenen auftreten, nicht immer erkannt, analysiert und gelöst werden. Eine Reihe verschiedener Ansätze fokussiert konkret auftretende Probleme in verschiedenen, kleinen Teilbereichen Digitaler Bibliotheken.

Um die unterschiedlichen Entitäten zu identifizieren, die für die Dokumentation von wissenschaftlichen Forschungsaktivitäten und -ergebnissen relevant sind, und um Relationen zwischen ihnen zu erkennen und zu analysieren, wurden Modelle wie z.B. der CERIF Standard[1] (Common European Research Information Format) von der Europäischen Kommission und euroCRIS entwickelt oder die PolicyGrid Ontology (vgl. Chorley et al. 2006). Diese Modelle können den kompletten Forschungsprozess darstellen, unter Berücksichtigung von Akteuren, Projekten, Organisationen, Publikationen, etc.

[1] http://www.eurocris.org/cerif/introduction

In einer globalen digitalen Umgebung spielt darüber hinaus die eindeutige Identifikation von Ressourcen und deren Langzeitverfügbarkeit eine gewichtige Rolle. Durch die Verwendung von URIs (Uniform Resource Identifier) besteht die Möglichkeit, Informationsressourcen wie Dokumente, Daten, Personen, etc. eindeutig adressierbar und referenzierbar zu machen. Ein URI kann jedoch auch Teilstrukturen wie einzelne Konzepte, Terme und Relationen eines Wissensorganisationssystems auszeichnen. Das Konzept der URIs sieht dabei jedoch vor, dass die jeweilige Adresse einer Ressource über ihren gesamten Lebenszyklus hinweg gleich bleibt. Persistente Identifikatoren wie Digital Object Identifier (DOI)[2] oder URNs verfolgen hingegen den Ansatz, Datensätze und Publikationen permanent und vom jeweiligen Speicherort unabhängig zu adressieren (Paskin 2008), was für die Langzeitverfügbarkeit eine wichtige Voraussetzung ist.

2.2 Struktureller Kontext

Auf strukturelle Ebene findet sich ein breites Feld an verschiedenen, in den jeweiligen Communities etablierten, Metadatenstandards wieder, wie beispielsweise das DDI-Format der Data Document Initiative[3] oder SDMX[4], um Primärdaten zu dokumentieren, oder die beiden bibliographischen Standards MARC[5] und Dublin Core[6]. Allgemeine Standards fokussieren in diesem Kontext eher auf die Entwicklung und Anwendung gemeinsamer Austauschformate für Metadaten über Harvestingschnittstellen wie OAI-PMH[7], als auf die Verständigung auf einem gemeinsamen Standard, was aber aufgrund der deutlich verschiedenen Anforderungen kaum anders zu bewerkstelligen ist. Durch bestehende Abbildungen (Mappings) zwischen den verschiedenen Formaten können allerding ähnliche Elemente aufeinander abgebildet werden. Die Forschungsfelder des Schema Mapping und des Metadatenaustausches sind bereits sehr weit bearbeitet und werden daher in diesem Paper nicht weiter ausführlich behandelt.

2.3 Semantischer Kontext

Auf semantischer Ebene finden sich viele Projekte und Initiativen, die sich mit dem Mapping von Terminologien und Vokabularen befassen. Hierbei lässt sich eine große Bandbreite an verschiedenen Techniken und Verfahren identifizieren, die jeweils für sich die semantische Heterogenität auf unterschiedliche Art behandeln (vgl. Zeng & Chan 2004). Ein semantisches Netzwerk zwischen verschiedenen Informationsquellen, beispielsweise zwischen verschiedenen Informationssammlungen, die wiederum mit verschiedenen Indexierungsvokabularen erschlossen sind, lässt sich durch Abbildungen der Terme des einen Vokabulars auf adäquate Terme des anderen aufbauen (vgl. Stempfhuber 2007). Auf diese Weise lässt sich eine integrierte Suche über verteilte und unterschiedlich sacherschlossene Informationsquellen umsetzen. Es existieren zahlreiche Methoden, um Vokabulare aufeinander abzubilden, die von

[2] http://www.doi.org
[3] http://www.ddialliance.org
[4] http://www.sdmx.org
[5] http://www.loc.gov/marc
[6] http://dublincore.org
[7] http://www.openarchives.org

intellektuellen Verfahren, z.B. Crosskonkordanzen (vgl. Mayr & Petras 2008a) bis hin zu automatisierten, statistischen oder deduktiven Verfahren reichen.

Die Behandlung semantischer Heterogenität gewinnt zunehmend an Komplexität, wenn grundverschiedene Informationsarten wie textuelle Daten (z.B. Publikationen) und Faktendaten (z.B. Umfragedaten) kombiniert werden, da bei letzterem semantische Informationen und damit relevante Inhalte nicht unbedingt in den Metadaten ausgedrückt sind, sondern in den Daten implizit enthalten sind. Bei Primärdaten finden darüber hinaus andere Indexierungsvokabulare wie Nomenklaturen oder Klassifikationen Anwendung als bei Publikationen, die vorwiegend mit Thesauri sacherschlossen werden. Eine Herausforderung stellt ein mögliches Mapping zwischen diesen verschiedenen Inhaltserschließungssystemen dar, da sie über unterschiedliche semantische Ausdrucksstärke verfügen, was sich z. B. in den innerhalb eines Systems zum Einsatz kommenden semantischen Relationen manifestiert. So verfügen Thesauri beispielsweise meist über hierarchische und assoziative Relationen, die jedoch für eine spezifische Erfassung von Primärdaten oft nicht ausreichen. Beide Ansätze sind durch ihre Anwendungsszenarien gerechtfertigt, sorgen jedoch für Abbildungsprobleme, wenn sie in einer integrierten Suche zur Anwendung kommen sollen: Bei Primärdaten finden sich die relevanten Informationen – die wissenschaftliche Intention eine bestimmte Frage oder einen Fragekomplex zu stellen – nur in wenigen Fällen in den jeweiligen Fragen oder Studienbeschreibungen wieder (vgl. Krause & Stempfhuber 2005). Meist sind diese Informationen nur implizit enthalten. In beiden Fällen kann eine solche Information nur schwer auf adäquate Einträge eines Thesaurus für Literatur abgebildet werden, da die dort zur Verfügung stehenden Relationen oft nicht ausdrucksstark genug sind.

Bei der Suche nach Literatur können Nutzer anhand der Titelliste relativ einfach entscheiden, ob eine Publikation für ihr jeweiliges Informationsbedürfnis relevant ist oder nicht. Im Kontext von Primärdaten findet eine solche Entscheidung eher auf Fragen- und Variablenebene oder durch eine Kombination verschiedener Merkmale (Erhebungsmethode, Variablen, Grundgesamtheit, etc.) statt. Daher ist eine höhere Präzision bei der Sacherschließung notwendig, um die Informationsbedürfnisse der Nutzer zu befriedigen.

Ontologien bieten die Möglichkeit, Daten mit ausdrucksstärkerer Semantik zu beschreiben. Da Ontologien bereits in Ansätzen auf architektonischer und organisatorischer Ebene (siehe Abschnitt 2.1) Anwendung finden, liegt es nahe sie auch auf semantischer Ebene zu verwenden. Insbesondere die Entwicklung von SKOS (Simple Knowledge Organization System)[8] dient dabei als Türöffner, traditionelle Wissensstrukturen wie Thesauri und Klassifikationen für das Semantic Web nutzbar und verarbeitbar zu machen und somit zumindest technisch eine Verlinkung mit Ontologien zu ermöglichen. Wie sie auf konzeptioneller Ebene effektiv und vor allem semantisch korrekt und adäquat mit Ontologien verlinkt werden können, bleibt noch zu untersuchen.

3 Modell für Text-Fakten-Integration in Digitalen Bibliotheken

Das im folgenden Abschnitt vorgestellte Modell behandelt die semantische Integration von heterogenen Informationstypen in Digitalen Bibliotheken. Es fokussiert dabei primär die

[8] http://www.w3.org/2004/02/skos

Behandlung der semantischen Heterogenität und kann einige der bereits dargestellten Problemstellungen und Herausforderungen lösen, indem diese unter Zuhilfenahme bereits existierender Entwicklungen und Technologien in einen gemeinsamen Kontext gesetzt werden. Es wird nicht nur die Semantik verschiedener Datentypen (z.B. von Umfragedaten oder Publikationen) berücksichtigt, sondern auch die zur Verlinkung dieser Datentypen untereinander benötigten semantischen Relationen, wobei der gesamte Forschungsprozess abgedeckt werden kann. Das Modell setzt sich aus drei Ebenen zusammen, die aufeinander aufbauen; dabei behandelt jede der drei Ebenen ein eigenes semantisches Modellierungsproblem im Speziellen (siehe Abbildung 1).

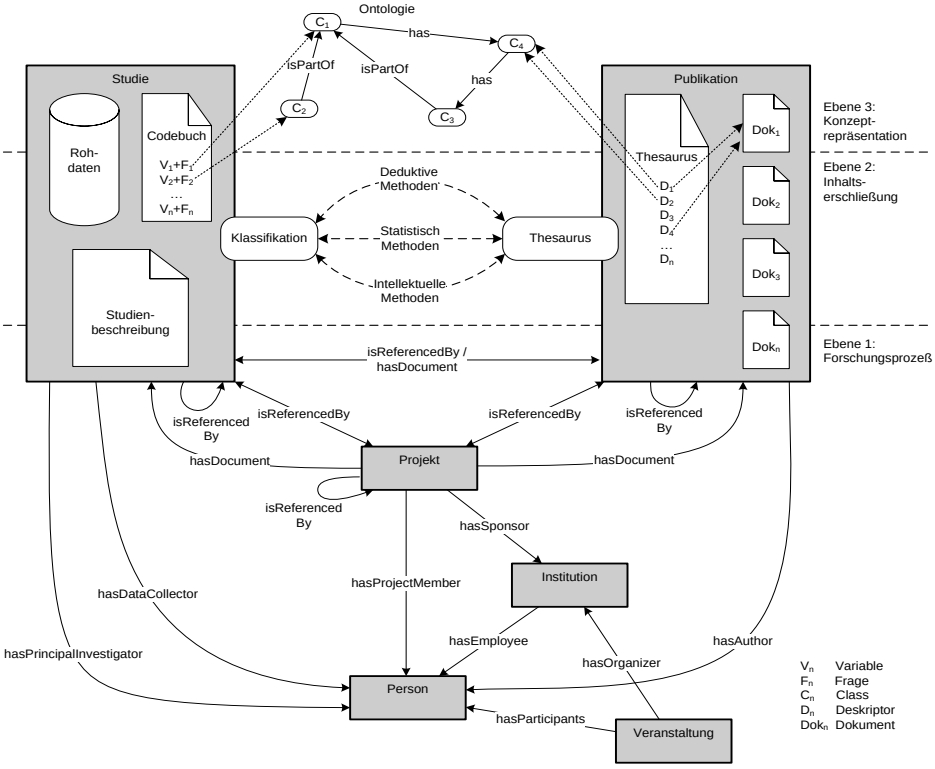

Abbildung 1: Modell für Text-Fakten-Integration

In den folgenden Abschnitten werden die drei Ebenen (Forschungsprozess, Inhaltserschließung und Konzeptrepräsentation) beschrieben. Dabei wird zusätzlich die Verbindung zum jeweiligen Forschungskontext gezogen und verdeutlicht.

3.1 Ebene 1: Forschungsprozess

Die erste Ebene (siehe Abbildung 2) reflektiert den kompletten wissenschaftlichen Forschungsprozess und zeichnet die Beziehungen zwischen allen beteiligten Entitäten (z.B. Per-

sonen, Institutionen, Forschungsprogramme, Projekte, Ergebnisse, Patente, etc.) aus. Bei einer Modellierung dieser Relationen mittels RDF (Resource Description Framework)[9] lassen sich auch komplexere Relationen und Beziehungen darstellen. Darüber hinaus repräsentiert die Ebene den Kontext, in dem Forschung durchgeführt wird und Ergebnisse produziert werden. Sie basiert auf etablierten Modellen wie dem CERIF Standard oder der PolicyGrid Ontologie (vgl. Chorley et al. 2006). Die Relationen dieser Ebene ermöglichen deduktive Prozesse innerhalb des Forschungsraumes, z.B. über die Autorenschaft von Forschungsergebnissen, die Verlinkung von Ergebnissen zu Forschungsprojekten oder die Verlinkung von verschiedenen Projekten zu Forschungsprogrammen. Anwendung finden diese Relationen beispielsweise beim Blättern durch verwandte Informationen. Sie bilden den Kern eines Forschungsinformationssystems, auf dem die anderen Ebenen basieren.

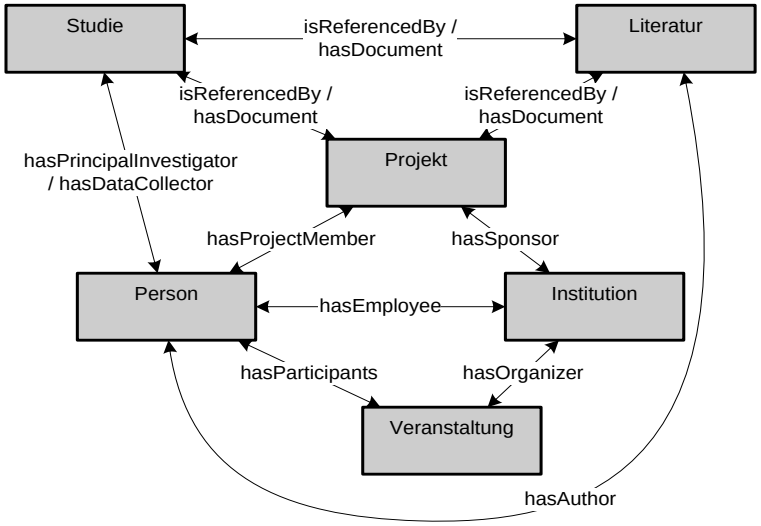

Abbildung 2: Forschungsprozess (Ebene 1)

Durch die Semantik, die diese Ebene implizit enthält, kann die Vagheit beim Retrieval reduziert werden, in dem die Ergebnisse bereits weiterführende (Hintergrund-)Informationen zu einzelnen Treffern enthalten können bzw. auf diese verweisen können. Neben der Auszeichnung aller Entitäten und deren Rollen (z.B. Personen in verschiedenen Rollen wie Autor, Forscher, Projektmanager, etc.) mit eindeutigen persistenten Identifikatoren können Informationen generiert werden, die normalerweise nicht in einer einzelnen Entität vermerkt sind, wie beispielsweise die strategischen Ziele eines Förderprogrammes. Diese relevanten Informationen können die Suchstrategien der Nutzer unterstützen.

[9] http://www.w3.org/rdf

3.2 Ebene 2: Inhaltserschließung

Diese Ebene (siehe Abbildung 3) behandelt die semantischen Informationen, die direkt in den Daten bzw. den Metadaten enthalten sind (also die Inhaltserschließung durch Schlagwörter oder Notationen aus Klassifikationen). Der Heterogenität zwischen unterschiedlichen Vokabularen, die in verschiedenen Informationssammlungen und für verschiedene Informationsarten verwendet werden (z.B. Klassifikationen und Nomenklaturen für Primärdaten oder Thesauri für Publikationen), wird begegnet, indem diese Vokabulare aufeinander abgebildet werden.

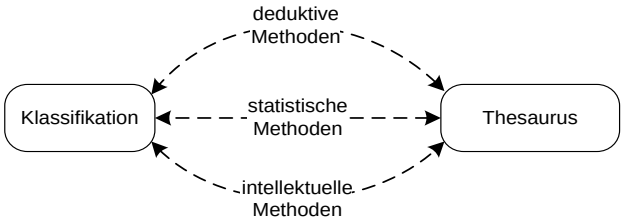

Abbildung 3: Inhaltserschließung (Ebene 2)

Es existiert eine Reihe von Ansätzen, die Behandlung der semantischen Heterogenität zu unterstützen. Dazu zählen intellektuelle Verfahren wie z.B. bilaterale Crosskonkordanzen, statistische und deduktive Verfahren, die den Recall beim Information Retrieval erhöhen (vgl. Krause 2004). Diese Verfahren können Nutzer besonders bei der Transformation von Suchtermen unterstützen, indem im Hintergrund die vom Nutzer eingegebenen Terme automatisch in adäquate Terme anderer Vokabulare transformiert werden, mit denen andere Informationsbestände erschlossen wurden. Ein Umformulieren und Erlernen anderer Vokabulare von Nutzerseite kann damit vermieden werden.

3.3 Ebene 3: Konzeptrepräsentation

Die dritte Ebene (siehe Abbildung 4) behandelt spezifische Unterschiede in der semantischen Ausdrucksstärke zwischen Thesauri, Klassifikationen, Codebüchern, etc., indem semantisch implizite Informationen innerhalb der Primärdaten (z.B. die wissenschaftliche Intention hinter einer Frage) auf die weniger ausdrucksstarken Terme eines Thesaurus abgebildet werden, um z.B. Publikationen zu indexieren.

Das Problem dieser unterschiedlichen Ausdrucksstärke wird erkennbar, wenn Suchbegriffe zu einem hohen Recall führen, da sie bei einer Vielzahl von Studien in Studienbeschreibung, Frageformulierung oder Variablenlabel verwendet wurden. Oft zeigt erst eine genaue Analyse der Studie, ob diese relevant für das ursprüngliche Informationsbedürfnis ist, da z. B. erst von einer Frageformulierung, die den Suchbegriff enthält, auf die sozialwissenschaftliche Intention der Frage geschlossen werden muss. Mittels Ontologien können solche konkreten und inhaltlich komplexeren Aspekte einer Studie modelliert werden und als Verlinkung zwischen den Termen eines Thesaurus auf der einen Seite und den komplexen Strukturen auf Studienseite dienen. Durch die Entwicklung von SKOS ist zumindest auf technischer Seite eine Umset-

zung dieses Konzepts möglich, inwieweit solch ein Mapping zwischen Ontologien und Termen eines Thesaurus konzeptionell und inhaltlich realisierbar ist, bleibt zu prüfen.

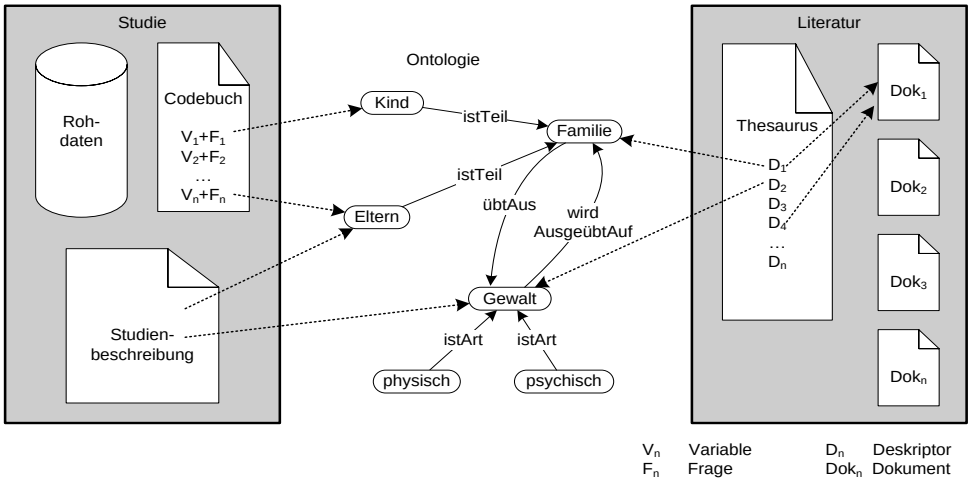

Abbildung 4: Konzeptrepräsentation (Ebene 3)

4 Zusammenfassung und Ausblick

Das vorgestellte Modell kombiniert komplementäre und parallel entwickelte Ansätze der Wissensorganisation und des Information Retrieval im Kontext von Digitalen Bibliotheken. Dabei wird die zu behandelnde Heterogenität sowohl auf struktureller als auch auf semantischer Ebene berücksichtigt. Eine integrierte Sicht, die auf bestehende Inhaltserschließungssysteme aufbaut und diese mit Ontologien kombiniert, reduziert den Aufwand ganze Domänen oder Disziplinen aufwändig zu modellieren, da nur konkrete Teilaspekte modelliert werden müssen, bei denen Thesauri nicht ausdrucksstark genug sind, um ein adäquates Retrieval über Publikationen und Primärdaten zu ermöglichen. Eingebettet in den Forschungsprozess werden durch das Modell zusätzlich semantische Relationen zu verwandten und assoziierten Ressourcen und Informationsarten aufgezeigt. Auch wenn der Blickwinkel auf die verschiedenen angewandten Ansätze durch deren Kombination untereinander deutlich erweitert wird, geht der Fokus auf die Detailebene dennoch nicht verloren.

Als Anwendungs- und Testszenario gleichermaßen dient für diese semantische Integration der GESIS-Datenbestandskatalog[10], der in die Suche des sozialwissenschaftlichen Fachportals sowiport.de[11] integriert werden soll. sowiport.de enthält bereits über 2,5 Millionen Literaturnachweise, Projektinformationen, Institutionsprofile, etc. Erste Evaluationsergebnisse bzgl. der zweiten Ebene des Modells, der Inhaltserschließung, belegen, dass der Recall relevanter Informationen deutlich verbessert wird (vgl. Mayr & Petras 2008b). Die erste und dritte Ebe-

[10] http://www.gesis.org/dienstleistungen/daten/recherche-datenzugang/datenbestandskatalog/
[11] http://www.sowiport.de

ne konnte noch nicht in diesem Umfang evaluiert werden, da speziell für ein integriertes Retrieval, wie es die dritte Ebene der Konzeptrepräsentation vorsieht, die technischen und konzeptuellen Umsetzungsmöglichkeiten genauer erforscht werden müssen. Generell besteht im Bereich des Information Retrieval im semantischen Web noch ein hoher Forschungsbedarf (vgl. Scheir et al. 2007; Finin et al. 2005). Zukünftige Arbeiten werden den Fokus auf die Implementierung und darauf folgende Evaluation der ersten und dritten Ebene des Modells setzen.

Obwohl das Modell im Kontext der Sozialwissenschaften entwickelt wurde, ist es nicht auf diese Disziplin beschränkt. Eine Anwendung in anderen Disziplinen und Domänen ist möglich, da bereits existierende und etablierte Wissensorganisationssysteme innerhalb des Modells genutzt werden können.

Literatur

ARL (2006): 'To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering', ARL/NSF Workshop on Long-Term Stewardship of Digital Data Collections, 2006.

Candela, L; Castelli, D.; Pagano, P. (2006): 'A Reference Architecture for Digital Library Systems', in: ERCIM News, Special theme: European Digital Libraries, 2006, No. 66.

Chorley, A.; Edwards, P.; Hielkema, F.; Philip, L.; Farrington, J. (2008): 'Developing Ontologies to Support eSocial Science: The PolicyGrid Experience', in Proceedings of the 4th International Conference on e-Social Science, Manchester, 2008.

DELOS (2005a): 'D5.3.1: Semantic Interoperability in Digital Library Systems', The DELOS Network of Excellence on Digital Libraries, 2005.

DELOS (2005b): 'The DELOS Network of Excellence on Digital Libraries: Recommendations and Observations for a European Digital Library (EDL)', 4th DELOS Brainstorming Workshop on Digital Libraries, December 2005.

Finin, T.; Mayfield, J.; Joshi, A.; Cost, R.; Fink, C. (2005): 'Information Retrieval and the Semantic Web', 38th Annual Hawaii Internation Conference on System Sciences. Waikoloa, Hawaii.

Goble, C., Corcho, O., Alper, P. and De Roure, D. (2006): 'e-science and the semantic web: A symbiotic relationship', in: Discovery Science 2006, Barcelona, Spain.

Gold, A. (2007): 'Cyberinfrastructure, Data and Libraries. Part 1 & 2', D-Lib Magazine, 2007, Volume 13, Number 9/10.

Gonçalves, M.; Fox, E.; Watson, L.; Kipp, N. (2004): 'Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries', ACM Trans. Inf. Syst, Volume 22, pp. 270—312.

Krause, J. (2004): 'Standardization, Heterogeneity and the Quality of Content Analysis: a key conflict of digital libraries and its solution', IFLA Journal: Official Journal of the International Federation of Library Associations and Institutions, 2004, No. 4, pp. 310-318.

Krause, J.; Stempfhuber, M. (2005): 'Nutzerseitige Integration sozialwissenschaftlicher Text- und Dateninformationen aus verteilten Quellen', in König, C.; et al. (eds.): Datenfusion und Datenintegration: 6. Wissenschaftliche Tagung, Bonn, 2005, pp. 141-158.

Mayr, P.; Petras, V. (2008a): 'Building a terminology network for search: the KoMoHe project', in: Greenberg, J.; Klas, W. (eds.): International Conference on Dublin Core and Metadata Applications (DC 2008). Berlin. pp. 177-182.

Mayr, P.; Petras, V. (2008b): 'Cross-concordances: terminology mapping and its effectiveness for information retrieval', IFLA World Library and Information Congress, 2008.

Paskin, N. (2008): 'Digital Object Identifier (DOI) System', in: Encyclopedia of Library and Information Sciences, 3rd Edition (to appear).

Poll, R. (2004): 'Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung. Teil 1: Informationsverhalten und Informationsbedarf der Wissenschaft', in: ZfBB, 2004, No. 51, pp. 59-75.

Scheir, P.; Pammer, V.; Lindstaedt, S. (2007): 'Information Retrieval on the Semantic Web – Does it exist?', in: Proceedings of Lernen-Wissen-Adaption, Germany, pp. 252-257.

Stempfhuber, M. (2007): Heterogeneity and Information Fusion Driven by User Needs. S. 153 - 170. In: Raghavan, K. S. (2007): IKONE 2007: International Conference on the Future of Knowledge Organization in the Networked Environment; 3 - 5 September 2007, Bangalore, India. Indian Statistical Institute Platinum Jubilee Conference Series.

Sure, Y.; Studer, R. (2005): 'Semantic Web Technologies for Digital Libraries', Library Management, Special Issue: Semantic Web, 2005, 26 (4/5), pp. 190-195.

Svensson, L. (2007): 'National Libraries and the Semantic Web: Requirements and Applications', in Prasad, A.R.D.; Madalli, D. (eds.): International Conference on Semantic Web and Digital Libraries, Bangalore, 2007.

Zeng, M.; Chan, L. (2004): 'Trends and issues in establishing interoperability among knowledge organization systems', Journal of the American Society for Information Science and Technology, 55(3), pp. 377-395.